Exhibit 1

Dynamic Voltage and Frequency Scaling Circuits with Two Supply Voltages

Wayne H. Cheng and Bevan M. Baas

ECE Department, University of California, Davis

***Abstract*— This paper presents circuits that enable dynamic voltage and frequency scaling (DVFS) for fine-grained chip multi-processors to reduce both dynamic and leakage power dissipation. Each processor can run on either a high voltage or low voltage power supply, or disconnect from both. Switching between power supplies is performed dynamically, where scaling decisions are based on each processor's workload, allowing for reduced power consumption without a significant impact on performance. Tradeoffs in performance versus circuit area and supply noise are examined. The DVFS circuits are designed in a wrapper around each individual processor, resulting in a 12% area overhead. DVFS operation utilizing supply voltages of 1.3 V and 0.8 V on a nine-processor JPEG application reduces average energy consumption by 48% while reducing performance by only 8%.**

I. Introduction

Increasing levels of power dissipation in processors built with advanced CMOS fabrication technologies has made power minimization a key design requirement. One solution to decreasing power consumption is through the use of dynamic supply voltage and dynamic clock frequency scaling (DVFS) techniques [1].

In this paper, the design of the DVFS circuit is implemented on an AsAP (Asynchronous Array of simple Processors) many-core chip [2]. This architecture contains processors that are small, simple, easily replicable, and each located in their own clock domain. DVFS circuits and techniques are applied to each processor core such that each core is contained in its own independent clock and voltage domain.

II. Background

Lowering the supply voltage leads to a square reduction in dynamic power based on the dynamic power-voltage relationship: $P_{dyn} = aCVdd^2f$, where a is the switching probability or activity, C is the total load capacitance, Vdd is the supply voltage, and f is the clock frequency. Without altering the supply voltage, power can be reduced with frequency reduction, but the energy consumption per operation remains the same. Supply voltage reduction on the other hand, contributes directly to energy reduction, where the dynamic energy consumption of a gate is a direct function of the supply voltage: $E = CVdd^2$.

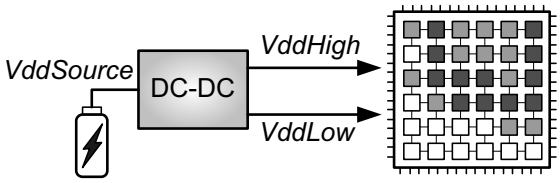

Fig. 1. Dynamic voltage and frequency scaling for a chip-multiprocessor

Leakage power is reduced as well with reduced supply voltages under normal circumstances. This is true for both sub-threshold leakage ($P_{sub\_leakage} \propto 1 - e^{Vdd}$) and gate leakage ($P_{gate\_leakage} \propto Vdd^2/e^{Vdd}$). DVFS becomes increasingly important as leakage power becomes a dominant contribution to power consumption in very deep-submicron CMOS technologies [3]. Benefits of DVFS also include counteracting process variations and thermal effects [4]. Slower parts of the chip can be sped up with higher voltages, and hotter parts can be cooled with lower voltages.

Reduction in supply voltage results in an increased gate propagation delay (t_d) [5]: $t_d \approx CVdd/(Vdd-V_t)^\alpha$, where V_t is the threshold voltage, and α is the velocity saturation index (≈ 1 in nanometer regimes). To guarantee correct operation of a synchronous system, the frequency must normally be scaled along with the voltage. The performance overhead of frequency and voltage scaling can be mitigated in a multi-processor architecture by taking advantage of the variation in workloads across the processor array. Processors can operate at a higher voltage during periods of high workloads, and at lower voltages during periods of low workloads to minimize energy dissipation.

III. Motivation and Implementation

Most DVFS implementations apply only a single DVFS controller to an entire chip using an on-chip or off-chip DC-DC converter. A fine-grain DVFS implementation can increase the effectiveness of DVFS by tuning the supply voltage to individual parts of the chip. However, it is impractical to design efficient DC-DC converters for many voltage domains on a single chip because the design of efficient converters requires large inductors and capacitors. A promising compromise is to supply discrete voltages to the chip, and have individual blocks switch between these voltages. By employing a voltage dithering method [6] to switch between discrete voltages, the

978-1-4244-1684-4/08/$25.00 ©2008 IEEE

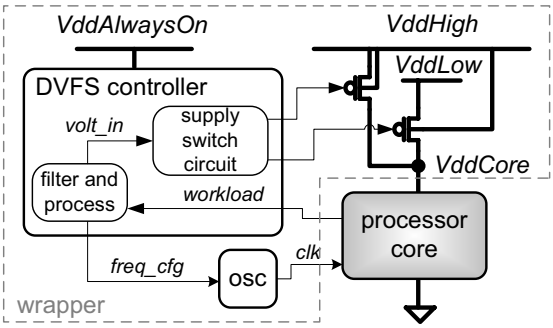

Fig. 2. Dynamic voltage and frequency scaling with two voltage supplies

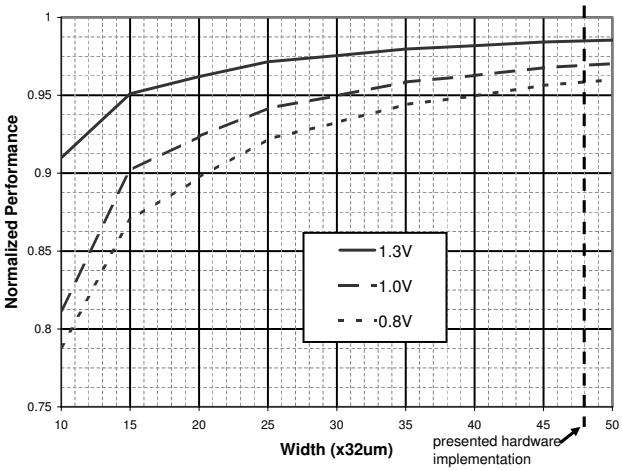

Fig. 3. Power gate transistor width versus processor performance

performance overhead of quantization can be reduced. Buffering of data is required to handle the quantization performance overhead. The diminishing maximum voltage associated with transistor scaling is a major limiting factor to voltage scaling [3]. As the maximum allowable voltage gets smaller, the advantages of having more than two discrete voltages diminish with the additional power, area, and supply switching delay overhead. Dithering between two discrete voltage levels has been experimented on an accumulator circuit [7] with promising results.

Figure 1 provides a concept diagram of DVFS on an AsAP multicore chip where two voltages are provided to the chip from an off-chip DC-DC converter. Each processor can choose to connect to either voltage supply, or none at all. Figure 2 shows DVFS with two voltage supplies using PMOS power gates [8]. The PMOS gates in the figure may represent many PMOS gates in parallel.

IV. Power Gate Design and Placement

Current flowing through power gate transistors results in a voltage drop that negatively impacts performance. The amount of voltage drop, V_{PG}, is related to the dimensions of the power gates: $V_{PG} = I_{PG}R_{PG}$, where I_{PG} is the current through the power gates, $R_{PG} \propto L/W$, and L and W are the length and width of the power gate transistors respectively. This voltage drop causes an increase in the power gate's delay [5]: $t_d \approx C\,Vdd/(Vdd - V_{PG} - V_t)^{\alpha}$. Voltage drop can be reduced by making W/L as large as possible, which can be accomplished by adding power gates in parallel. Additional decoupling capacitors, which act as low pass filters, can also be added to reduce power grid voltage fluctuations.

To accurately measure the performance loss associated with the power gates, a precise current profile from the processor core is first obtained with SPICE simulations. This current waveform is then used to create the voltage drop across the power gates, and the resulting increase in delay can be measured. In a 65 nm technology, the relationship between power gate width and performance at 25 °C is shown in Fig. 3.

The wrapper design is shown in Fig. 2. Designing the DVFS circuit as a wrapper allows for a straightforward substitution of the processor with another logic unit. The power gates are positioned in a vertical fashion so that the power gates are aligned with the vertical power stripes, as shown in Fig. 6. Power is distributed across the processor through horizontal power stripes.

The DVFS controller in Fig. 2 contains logic to estimate the workload (indicated by *workload*) by the FIFO utilization and stall duration of the processor. The workload information is filtered and processed with configurable FIR and IIR filters. Frequency scaling is performed by incrementing or decrementing the clock frequency based on the workload information. A range of allowable frequencies is assigned for each voltage setting, where the settings of *volt_in* are mapped to settings of *freq_cfg*. Therefore, voltage scaling is performed automatically depending on the frequency setting of the processor.

V. Robust Dynamic Run-Time Voltage Switching

Switching between voltage supplies during run-time results in supply grid noise, possible shorting between supplies, and possible corruption of stored data. These issues are addressed in the run-time voltage switch design in Fig. 4, and the corresponding timing diagram in Fig. 5. Following a request for a voltage switch (where the signal *volt_in* changes), correct operation of the processor is guaranteed by sending a stall request before the actual switching of voltage. Stalling prevents processor operation during the period when the voltage supply is not completely connected, so that stored data are preserved within the internal circuits. When stalling is finished, a confirmation signal (*stall_done*) is transmitted back to the supply switch circuit. Shorting between power supplies is prevented by first shutting off power gates for both supplies with the *force_off* signal. A configurable amount of delay is provided between the switching of power supplies by the "variable delay mechanism" which is implemented with a delay chain and a multiplexer. Upon completion of the delay, the *force_off* signal is released by the *delay_done* signal, and the power gates are then switched on to the new power supply. Finally, the stall signal is released after the new voltage supply is fully connected.

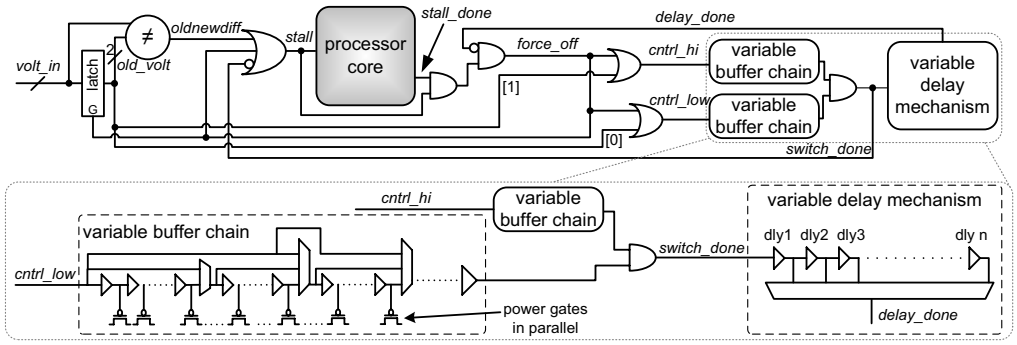

Fig. 4. Configurable dynamic run-time voltage supply switching circuit for the DVFS controller shown in Fig. 2

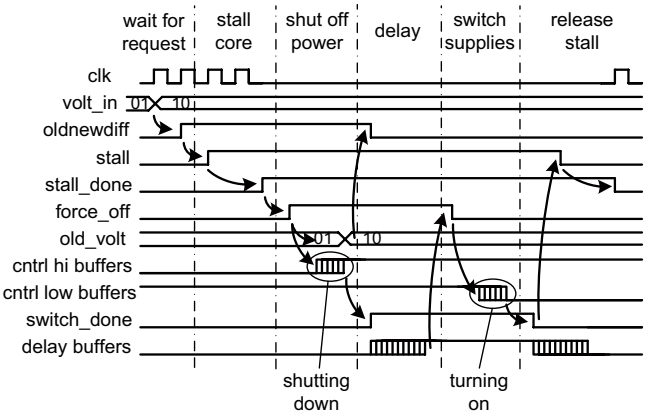

Fig. 5. Timing diagram for the supply switching controller changing the supply voltage from Vdd_{High} to Vdd_{Low}

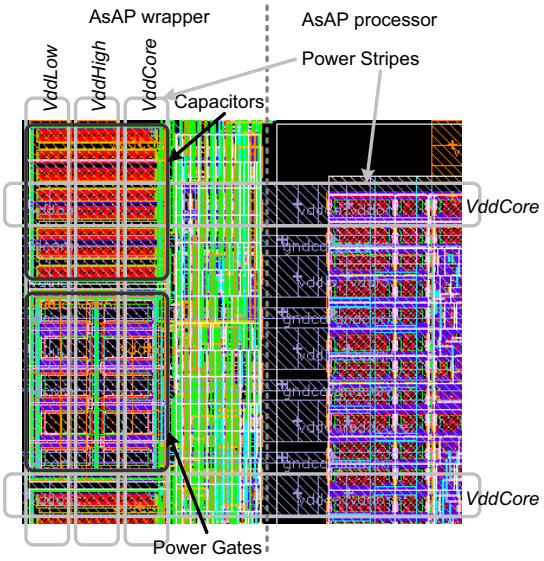

Fig. 6. Zoomed in view of the DVFS wrapper and the AsAP processor core

Noise on the power grid occurs when switching a processor from one voltage supply to the other. A processor switching to a higher voltage supply grid will cause a droop in the voltage of the higher voltage grid, and switching to a lower voltage supply will cause an upward surge in the voltage of the lower voltage grid. A droop in voltage will momentarily increase the gate delay and may cause circuit failures. To limit supply grid noise, the "variable buffer chain" allows for a configurable rate of shutting off and turning on of the power gates. The power gates can be configured to switch gradually by allowing the control signal to propagate through the buffer chain. If performance is critical, the voltage switching time can be reduced by configuring the power gates to turn off and on simultaneously.

VI. Results

The DVFS circuits were implemented in each AsAP processor in 65 nm CMOS technology. Details of the power gate layout supplying power to the processor's supply, *VddCore,* are shown in Fig. 6. There are a total of 48 power gates per processor, each with a width of 32 μm, for each power supply. Decoupling capacitors shown connect to *VddCore.*

The total area of the DVFS wrapper design occupies approximately 12% of an AsAP processor core area. About 66% of the DVFS wrapper area is devoted to the power gates and decoupling capacitors. The maximum power consumption of the DVFS wrapper is approximately 4% of the typical power of an AsAP processor.

To test the effectiveness of the DVFS circuit, the behavior of the oscillator over different frequency settings was ported from SPICE to a logic simulator, and the operating times on each voltage supply was tabulated. Using the following formula, an estimate of the relative energy delay product, EDP_{rel}, was calculated:

$$EDP_{rel} = \left(\frac{\beta\, Vdd_{Low}^2 + (1-\beta)\, Vdd_{High}^2}{Vdd_{High}^2} \right) \left(\frac{t_{dvfs}}{t_{orig}} \right) \quad (1)$$

where β is the fraction of time operating on the lower voltage supply, t_{dvfs} is the total run time with DVFS, and t_{orig} is the total run time without DVFS.

The behavior of a nine-processor JPEG application was analyzed with different configuration settings. Simulations were performed with voltage supplies of $Vdd_{High} = 1.3$ V and $Vdd_{Low} = 0.8$ V (chosen as the maximum and minimum volt-

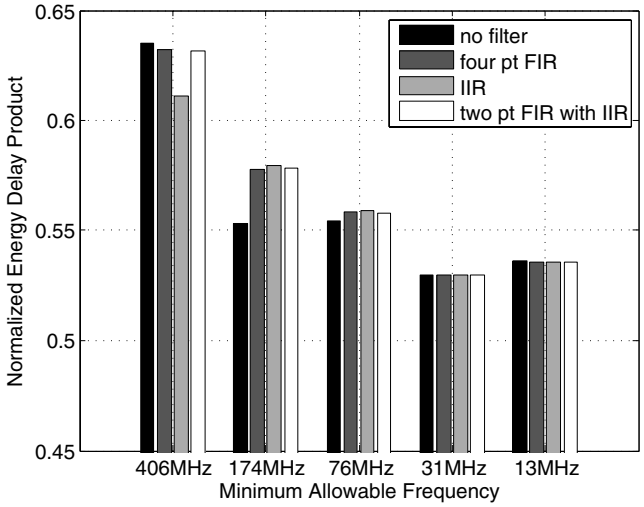

Fig. 7. Relative energy delay product with DVFS compared to a non-DVFS design on a 9 processor JPEG application, with a maximum frequency of 1.05GHz

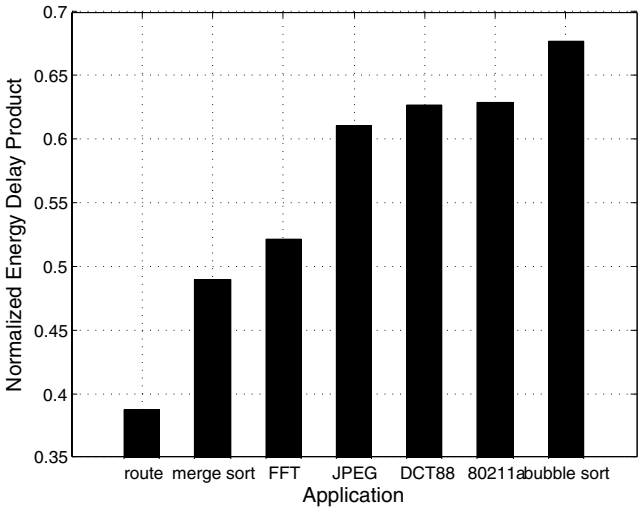

Fig. 8. Relative energy delay product with DVFS compared to a non-DVFS design for various applications, with a frequency range of 150MHz to 1.05GHz

age within technology specifications). Figure 7 illustrates the effects of different filter and allowable frequency settings on the relative energy delay product. Filter settings affect results, and are application and configuration specific. Decreasing the minimum allowable frequency on the lower voltage supply will require more frequency increments to switch to the higher supply, and therefore increase the operating time on the lower supply. As the minimum frequency is decreased (from left to right in Fig. 7), the resulting energy delay product decreases, where the energy savings outweigh the performance overhead. This is no longer true when the minimum frequency is too low (at 13MHz in the figure) and the performance overhead outweighs the energy savings. Running with DVFS resulted in an average of 52% of the original energy consumption, with an 8% performance overhead. The average of the results in Fig. 7 is 0.56.

The effectiveness of the DVFS circuit (with the same voltage configuration settings) is also analyzed with different applications in Fig. 8. The behavior of each processor's workload (indicated by the FIFO utilization and stall duration) directly effects the performance overhead and *EDP*. In the merge sort application, each processor's workload is constant (either constantly large or small), and DVFS results in a low performance overhead and therefore a low *EDP*. Conversely, in the bubble sort application, each processor's workload oscillates between large and small, and DVFS results in a larger performance overhead and therefore a larger *EDP*. The average of the results in Fig. 8 is also 0.56.

VII. Conclusion

Dynamic voltage and frequency scaling was accomplished with two discrete supply voltages. Power gates were designed and sized to reduce performance loss. A robust dynamic run-time voltage switching circuit was developed that avoids shorting between supplies and excessive power grid noise.

A DVFS circuit is designed in a wrapper for each AsAP processor. On a 9-processor JPEG application, running with DVFS resulted in an average of 52% of the original energy consumption, with an 8% performance reduction and a relative *EDP* of 56%.

Acknowledgment

The authors gratefully acknowledge fabrication by ST Microelectronics; support from Intel, UC Micro, NSF Grant 430090 and CAREER award 546907, SRC GRC Grant 1598, Intellasys, S Machines, and Uniquify; and thank J.-P. Schoellkopf, K. Torki, R. Krishnamurthy, and M. Anders. T. Mohsenin, D. Truong, Z. Yu and other VCL members contributed to the AsAP many-core chip and applications.

References

[1] S. Borkar, "Low power design challenges for the decade," in *Proceedings of the 2001 conference on Asia South Pacific design automation*, 2001.

[2] Z. Yu, M. Meeuwsen, R. Apperson, O. Sattari, M. Lai, J. Webb, E. Work, T. Mohsenin, M. Singh, and B. Baas, "An asynchronous array of simple processors for DSP applications," in *IEEE International Solid-State Circuits Conference, (ISSCC '06)*, Feb. 2006.

[3] "International technology roadmap for semicondutors," in *http://www.itrs.net/reports.html*, 2006.

[4] Y. Liu, H. Yang, R. Dick, H. Wang, and L. Shang, "Thermal vs energy optimization for DVFS-enabled processors in embedded systems," in *Proceedings of the 8th International Symposium on Quality Electronic Design*, 2007.

[5] M. Anis, S. Areibi, M. Mahmoud, and M. Elmasry, "Dynamic and leakage power reduction in MTCMOS circuits using an automated efficient gate clustering technique," in *39th Design Automation Conference*, 2002.

[6] V. Gutnik and A. Chandrakasan, "Embedded power supply for low-power DSP," in *IEEE Transactions on VLSI Systems, Vol. 5, No. 4*, Dec. 1997.

[7] B. Calhoun and A. Chandrakasan, "Ultra-dynamic voltage scaling (UDVS) using sub-threshold operation and local voltage dithering," in *IEEE Journal of Solid-State Circuits, Vol. 41, No. 1*, Jan. 2006.

[8] Wayne H. Cheng, "Approaches and designs of dynamic voltage and frequency scaling," M.S. thesis, University of California, Davis, CA, USA, 2008.